Structure
Ways & Means

Accurate NMR Structures Through Minimization of an Extended Hybrid Energy

Michael Nilges,[1,*] Aymeric Bernard,[1] Benjamin Bardiaux,[1] Thérèse Malliavin,[1] Michael Habeck,[2,3] and Wolfgang Rieping[4]
[1]Institut Pasteur, Département de Biologie Structurale et Chimie, Unité de Bio-informatique Structurale, CNRS URA 2185,
25–28 rue du docteur Roux, F–75015 Paris, France
[2]Max-Planck-Institute for Developmental Biology, 72076 Tübingen, Germany
[3]Max-Planck-Institute for Biological Cybernetics, 72076 Tübingen, Germany
[4]Department of Biochemistry, University of Cambridge, 80 Tennis Court Road, Cambridge CB2 1GA, UK
*Correspondence: nilges@pasteur.fr
DOI 10.1016/j.str.2008.07.008

SUMMARY

The use of generous distance bounds has been the hallmark of NMR structure determination. However, bounds necessitate the estimation of data quality before the calculation, reduce the information content, introduce human bias, and allow for major errors in the structures. Here, we propose a new rapid structure calculation scheme based on Bayesian analysis. The minimization of an extended energy function, including a new type of distance restraint and a term depending on the data quality, results in an estimation of the data quality in addition to coordinates. This allows for the determination of the optimal weight on the experimental information. The resulting structures are of better quality and closer to the X–ray crystal structure of the same molecule. With the new calculation approach, the analysis of discrepancies from the target distances becomes meaningful. The strategy may be useful in other applications—for example, in homology modeling.

INTRODUCTION

The key aims of a structure calculation are to estimate the coordinates and their uncertainty, and to provide a meaningful measure of the quality of the fit to the data. The calculation strategy has to take into account the uncertainties in the data, and also the fact that experimental data are rarely sufficient to determine the three-dimensional structure of a macro-molecule by themselves but need to be complemented with prior physical information. Structure calculation is typically a search for conformations that simultaneously have a low physical energy, $E_{phys}(X)$, and minimize a cost function, $E_{data}(X)$, that quantifies the discrepancy between a structural model X and the data. Several parameters have a critical influence on the result of a structure calculation; most importantly, the precise way uncertainties in the measurements are included into the cost function, $E_{data}(X)$, and the relative weight we put on the data in comparison to the physical energy.

The use of generous bounds for distances derived from NOE measurements has been virtually synonymous with NMR structure determination for nearly three decades. At first glance, bounds seem to represent a good way to incorporate the uncertainties in the distances into the calculation. However, the correct estimation of the bounds is a crucial step for the success of a structure calculation, and necessitates the knowledge of the uncertainty in the data. Very often, data points need to be modified individually to remove violations in the structures and obtain models with low energy. This ad hoc modification of the experimental data is unsatisfactory, since it introduces strong human bias into the structure calculation and since we do not have generally accepted rules for the derivation of bounds (and for the modification of individual bounds). Bounds also lead to loss of information content, and the width of the bounds has an obvious influence on the information content of the restraints (Nabuurs et al., 2003) and, hence, on the estimate of coordinate uncertainty. More importantly, this information loss can lead to errors in the derived structures that go unnoticed (Nabuurs et al., 2006).

In this paper, we show that, by exploiting the data to an optimal extent during the calculation and avoiding bounds, structures of better accuracy and quality can be obtained than with standard methods. The new elements are a distance restraint potential derived from the log-normal distribution; a new "joint" hybrid energy function containing an additional term depending directly on the data quality; and a minimization scheme for the joint hybrid energy function that adapts the weight on the data during the calculation to the estimated data consistency to avoid overfitting. The "log-harmonic" potential derived from the log-normal distribution has several useful properties for structure calculations from distance restraints, in particular if the data contain inconsistencies. We show that the structures produced with this approach are of better quality and closer to the respective X–ray crystal structures than those calculated with other restraint potentials. In the approach, we deliberately keep a minimal force field, E_{phys}, that describes only the geometrical properties of the molecule, and thus follow a different philosophy to the recently proposed refinement approach using ROSETTA (Qian et al., 2007).

PROCEDURES AND RESULTS

The new calculation strategy is based on the standard idea of minimizing a hybrid energy function. Usually, this energy has the form (Jack and Levitt, 1978; Kaptein et al., 1985; Brunger et al., 1993):

$$E_{hybrid}(X) = E_{phys}(X) + w_{data}E_{data}(X), \quad (1)$$

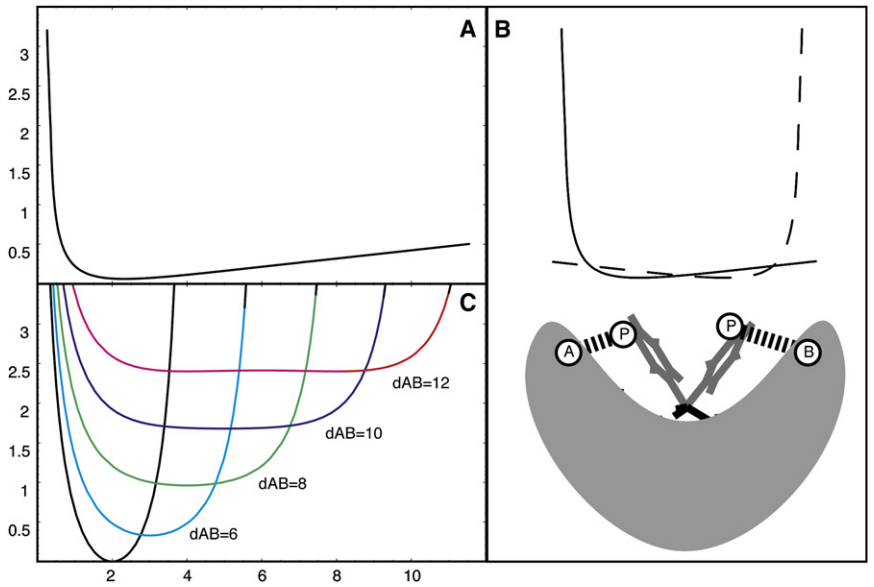

Figure 1. Log-Harmonic Potential

(A) Log-harmonic potential for a target distance of 2.5 Å.

(B) Illustration of inconsistent restraints to a proton P from two protons, A and B. Due to the flexibility of a side chain, a proton P shows an NOE to two protons A and B, despite a large distance between A and B. The resulting restraints are indicated.

(C) Effective potential for a proton P with distance restraints to two protons A and B. The target distances d_{AP} and d_{BP} are set to 2 Å, the distance d_{AB} between A and B is set to 4 (black line), 6 (cyan), 8 (green), 10 (blue), and 12 (red) Å, respectively.

where the force field, E_{phys}, compensates for the lack of data by imposing physical constraints on the structure, and $E_{data}(X)$ is the cost function that quantifies the disagreement between a structural model X and the data. The weight, w_{data}, controls the contribution of the data relative to the force field. In the standard approach, this term needs to be estimated by empirical means (Jack and Levitt, 1978; Brunger et al., 1990; Brunger and Nilges, 1993).

The new method modifies and extends this approach to estimate the weight, w_{data}. To this end, we combine three recent concepts into one rapid and efficient minimization protocol:

(1) We replace distance bounds by an error-tolerant potential with a single minimum, which we call the log–harmonic potential. The shape of this potential is derived from the log-normal distribution, which is a natural choice for distances and NOE volumes, and it models errors and inconsistencies in the data well (Rieping et al., 2005b). This potential has only one free parameter, its weight.

(2) We introduce an iterative automatic procedure suggested by Bayesian analysis (Habeck et al., 2006) to optimize the weight on the experimental data. This removes the one free parameter of the log-harmonic potential.

(3) The total energy of each structure is evaluated as the sum of three terms: the physical energy, E_{phys}, the restraint energy, E_{data}, and an additional term, E_σ, depending explicitly on the data quality, which is introduced by the Bayesian analysis (Habeck et al., 2006):

$$E_{joint}(X,\sigma) = E_{phys}(X) + w_{data}E_{data}(X) + E_\sigma. \quad (2)$$

The Log-Normal Distribution and the Log-Harmonic Potential

We recently showed that NOE intensities and derived distances follow a log–normal distribution (Rieping et al., 2005b)—that is, a normal (Gaussian) distribution in the logarithms of the data:

$$g(d_{obs}, d_{calc}(X)) = \frac{1}{Z(\sigma)}\exp\left[-\frac{1}{2\sigma^2}\chi^2(X)\right]. \quad (3)$$

Here, σ is the shape parameter of the distribution (similar to the standard deviation for the normal distribution), $Z(\sigma)$ is a normalization factor (Habeck et al., 2006), and χ^2 measures the discrepancies between the experimental data and the data calculated from the structure. For the lognormal distribution, we use the discrepancies in the logarithms of the data:

$$\chi^2(X) = \sum_i \log^2\left[\frac{d^i_{obs}}{d^i_{calc}(X)}\right]. \quad (4)$$

In contrast to the normal distribution, the log-normal distribution is restricted to positive values and is asymmetric around its median, $d_{calc}(X)$. Measurements are incorporated without bias in the sense that the probabilities of over- or underestimating the true value are both 1/2. We showed previously (Rieping et al., 2005b) that the use of this distribution in generating structures using the Inferential Structure Determination method (Rieping et al., 2005a) leads to structures of better quality than flat-bottom potentials.

The negative logarithm of this distribution represents the corresponding restraint potential (see Figure 1). If we identify, as before (Habeck et al., 2006) the restraint energy, E_{data}, with $\chi^2(X)/2$ and w_{data} with $1/\sigma^2$, we obtain the total weighted energy due to NOE derived distance restraints:

$$w_{data}E_{data} = \frac{1}{\beta}\frac{1}{2\sigma^2}\chi^2(X) \quad (5)$$

β is $1/k_BT$ and defines the energy scale; it is 1 if we measure the energy in units of k_BT. For clarity, it is therefore suppressed in the equation.

In contrast to flat-bottom potentials, the log-harmonic potential has one well-defined single minimum and has the opposite behavior: it is more restrictive for small deviations from the experimental distance, d_{obs}, but more tolerant to large violations (the asymptotic value of the slope is zero). The potential has the interesting property that inconsistent distance restraints to the same proton result in wider, softer potential wells and can result in multiple minima. This is illustrated in Figure 1C; the distance to

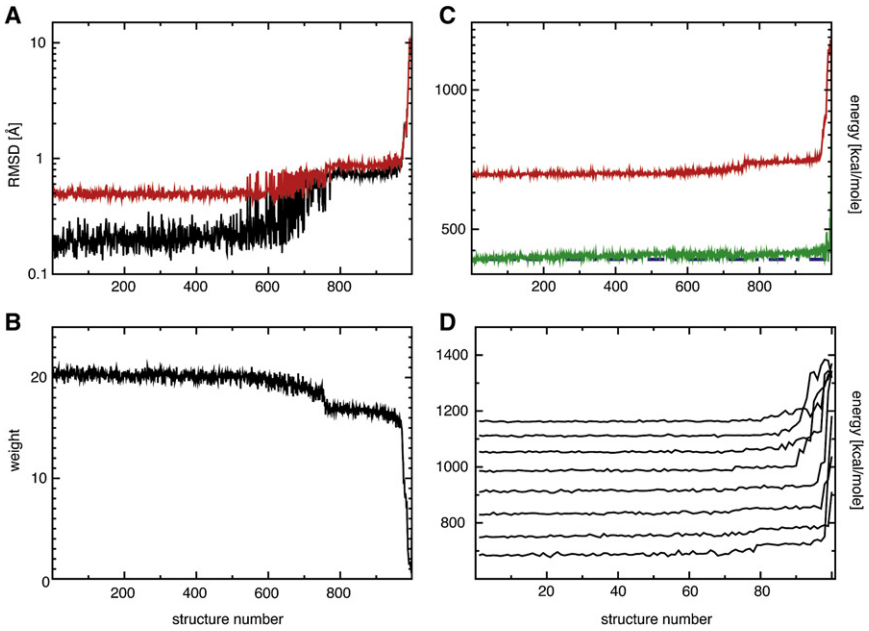

Figure 2. Structural Statistics for Ubiquitin
(A) RMS difference to the X–ray crystal structure 1UBQ (Vijay-Kumar et al., 1987) (red) and to the average structure (black), calculated from the 300 best structures.
(B) Weight ,w_{data}, calculated with Equation 7, for 1000 structures.
(C) Energy terms for 1000 structures of Ubiquitin: E_σ (red), $E_{data}(X)$ (blue, dot-dashed), and $E_{phys}(X)$ (green).
(D) Dependence of E_σ on data quality. Noise was added randomly with an amplitude ranging from 0.1 Å (bottom line) to 0.8 Å (top line).

a central proton from two exterior protons is restrained to 2 Å, and that the distance between the two exterior protons is 4, 6, 8, 10, and 12 Å. This type of inconsistency could occur if the central proton is on a mobile side-chain oscillating between two positions that are each close to one of the two exterior protons.

Minimizing the Joint Energy and Automated Determination of the Optimal Weight

The identification of w_{data} with $1/\sigma^2$ would directly give us the optimal weight if we knew σ, which depends on the quality of the data and on the correctness of the theory that relates molecular coordinates and the data. Both are a priori unknown, but a recent analysis solves this long-standing problem of optimally weighting experimental data (Habeck et al., 2006). This becomes possible by considering the two terms in the equation for the probability that depend on σ but have so far been neglected. The prior probability, π(σ), is required by Bayes's theorem to include prior knowledge about σ, and the factor $Z(\sigma)$ is necessary to normalize the error distribution. The additional term E_σ, depending on σ in Equation 2, is due to these two terms and their inclusion leads to the extension of the hybrid energy function, $E_{hybrid}(X)$ (Equation 1), to the joint energy function $E_{joint}(X,\sigma)$:

$$E_{joint}(X,\sigma) = E_{phys}(X) + \frac{1}{2\sigma^2}\chi^2 + \log\left[\frac{Z(\sigma)}{\pi(\sigma)}\right]. \quad (6)$$

The minimum of $E_{joint}(X,\sigma)$ defines the most probable structure, X_{max}, and the most probable error, σ_{max}. One can show (Habeck et al., 2006) that, for the log-normal model (Equation 3), the most probable error is a simple function of the coordinates, $\sigma_{max} = \sqrt{\chi^2(X_{max})/(n+1)}$, and that for the average weight we obtain simply

$$\langle w_{data} \rangle = \frac{n}{\chi^2(X)}, \quad (7)$$

where n is the number of data points. We use this estimate in our scheme to minimize $E_{joint}(X,\sigma)$.

How does one minimize $E_{joint}(X,\sigma)$ to obtain the most probable structure with the appropriate weight on the data? Within ISD (Rieping et al., 2005a), a probabilistic framework for structure determination, this weight is estimated along with all other unknown parameters by sampling over different values. Within the context of minimization of hybrid energy, we propose an iterative scheme: during the structure calculation, we iteratively update the current weight using Equation 7. In the simulated annealing protocol described further below, this update is performed during the final cooling phase every 1000 steps of dynamics.

With this procedure, each structure in an ensemble is calculated with its own weight, w_{data}, which is adapted to how well this structure fits the data. This has the startling result that the restraint energy does not depend on the structure (see Figure 2C) but in fact only on the number of data points. Our experience shows that the physical energy, E_{phys}, is rather uniform for all structures, since distortions are avoided for incorrectly folded structures due to the automated down weighting of $E_{data}(X)$. By itself, neither of these terms is therefore useful to distinguish converged from incorrect structures in the ensemble. The distinction is made by the term log $(Z(\sigma)/\pi(\sigma))$, which is monotonically increasing with σ (for the log-normal distribution, Equation 3, it is $\alpha\sigma^{n+1}$) (Habeck et al., 2006). For this reason, minimizing the restraint energy directly with respect to the weight does not automatically favor large values for σ, with the corresponding weight approaching 0.

We treat here only the log-normal distribution for distances; we stress, however, that the analysis is valid for data following other statistics (for example, coupling constants or residual dipolar couplings).

Structure Calculations with the Log-Harmonic Potential

Structures were calculated with an extended version of ARIA2/CNS (Rieping et al., 2007), using the ARIA-simulated annealing protocol with torsion angle dynamics and Cartesian dynamics (Linge et al., 2001), and a modified version of CNS (Brunger et al., 1998) that allows for the calculation of the optimal weight during the structure calculation. To speed up convergence, we used a soft-square potential for the high temperature phase with standard parameters (an asymptotic slope of 1.0 for violations larger than 0.5 Å, with the weight, w_{data}, set to 50). During

the second Cartesian cooling phase, the log-harmonic potential was used, and w_{data} was iteratively updated according to Equation 7. We used separate weights for NOE–derived distance restraints and for hydrogen bond distance restraints when the latter were used. The weight on the dihedral angle restraint term (which is not of χ^2 form—hence the automated weighting procedure cannot be used) was set to 5 (rather than the standard value of 200). To account for the slower convergence with the log-harmonic potential and the additional time needed to adjust the weight, we increased the number of steps to 2800 torsion angle dynamics steps in the high temperature phase, 2100 torsion angle dynamics steps in the torsion angle cooling phase, 20,000 steps in the first Cartesian dynamics cooling phase, and 20,000 in the second. The rest of the protocol (e.g., the variation of the force constant of the van der Waals interaction, variation of the temperature) remained unchanged. The time step was set to 45 fs in the torsion angle dynamics stages and to 5 fs in the Cartesian dynamics stages.

To demonstrate the method, we used the experimental data sets used in a previous study (Nilges et al., 2006) (iL4 [Powers et al., 1992], BPTI [Berndt et al., 1992], the PH domain of β Spectrin [Nilges et al., 1997], GB1 [Gronenborn et al., 1991], Ubiquitin [Cornilescu et al., 1998]), with the addition of iL8 (Clore et al., 1990). Data sets were derived from the reported distance bounds as described before (Nilges et al., 2006; see Supplemental Experimental Procedures available online). We calculated 100 structures for each data set and used the best 30 for analysis, with the exception of the noise-free data set for Ubiquitin, were we calculated 1000 structures.

As previously studied (Nilges et al., 2006), average structures were calculated with a modified version of the ''well-ordered'' script (Nilges et al., 1987), where atoms are not excluded from the fit but weighted iteratively to the inverse of the variance of the atom around the average structure. The original method (Nilges et al., 1987) would correspond to using only weights of 0 and 1. The weighting scheme used here originates from rigorous Bayesian analysis (Rieping, 2004) and is different from a recently published ad hoc weighting scheme (Damm and Carlson, 2006).

Structure and Data Quality

Figures 2A–C compare the RMS differences to the X–ray and the average structure with the weight, w_{data}, and the three energy terms in Equation 6 for the 1000 Ubiquitin structures calculated with the log-harmonic potential, using the recalibrated but otherwise unmodified distance restraints.

The structures were ordered with respect to total energy, $E_{joint}(X,\sigma)$. Almost all structures had converged to the same fold. Within the first half of the structures, there is little variation in the RMS differences. In the second half, both RMS differences increase; this rise is much more pronounced for the RMS difference from the average structure. The figure demonstrates that the energy, $E_{joint}(X,\sigma)$, correlates with the convergence to the average and reference structures and it is thus a good indicator of the quality of a structure.

Figure 2C shows the remarkable behavior of the individual energy terms themselves. The pseudo energy, $E_{data}(X)$, remains strictly constant; the physical energy, $E_{phys}(X)$, remains nearly constant, indicating that even in less well or unconverged structures, the covalent geometry is not distorted. In contrast, the new energy term, E_σ, changes significantly. After structure 550, this energy rises slightly, coinciding with the increased scatter and the beginning of the rise in the RMS difference to the average and X–ray crystal structures. This is followed by a plateau with slightly increased energy, from approximately structure 750 to 950, and a steep rise in energy for unconverged structures.

The constant value of the pseudo energy, $E_{data}(X)$, is a consequence of the estimate of the weight by Equation 7. The value obtained for σ, and therefore the weight, is quite similar to what one would use as a trivial estimate (Press et al., 1986) in the absence of any information on the data quality. As noted (Press et al., 1986), this trivial estimate removes the ability to judge the quality of the fit in least-squares analysis. However, the decisive difference between just using Equation 7 as an ad hoc estimate of the weight and our approach is that the latter provides a measure of the quality in the form of the joint energy, $E_{joint}(X,\sigma)$.

This estimate of the data quality is obtained as a result of the structure calculation, in contrast to basically all other methods, which need it as input to the structure calculation. The quality of the data sets in the present study, as analyzed by the final value of the RMS difference and consequently the weight, is fairly similar (it varies roughly by a factor of three). The weight is much more important with a single minimum potential than with a flat-bottom potential and has a significant influence on the quality of the obtained structures (Nilges et al., 2006). Adapting the weight is also important in order to obtain an unbiased estimate of the quality of the data from the calculation. The calculation of the weight during the structure calculation is straightforward and does not cost much computer time; we use a standard starting value and the weight is then iteratively adapted.

As demonstrated in Figure 2D, the method also works for more noisy data sets, where the experimental data set was made increasingly noisy by adding random noise from 0.1 to 0.8 Å to the estimated distance, d_{obs}. Only the new energy term, E_σ, is shown, since the other energy terms are completely (for $E_{data}(X)$) or nearly (for $E_{phys}(X)$) insensitive to the noise in the data. Whereas complete cross-validation (Brunger et al., 1993) was rather insensitive to the noise in the data (Nilges et al., 2006), one can clearly see the rise in the energy term, E_σ, with the level of noise.

Figure 3 shows a similar analysis for Ubiquitin in a sequence dependent manner. The RMSD from the average structure (Figure 3A) indicates that the calculation with the log-harmonic potential leads to rather tight ensembles, with an overall RMS difference to the average structure of 0.2 Å for the best 300 structures. Interestingly, the regions that have been reported to show dynamics behavior (see Clore and Schwieters, 2004) are considerably less ordered, in particular the loop near residue 10 (see Figure 4A and Figure S2 for a comparison of order parameters calculated from the ensemble and experiment).

An analysis of logarithmic discrepancies from the target distances (Figures 3B and 3C) shows that these discrepancies are distributed over the whole sequence in a more or less uniform manner (i.e., there are no regions in the protein where they cluster). We note that around 5% of the data are expected to deviate by more than two σ if they follow a log-normal distribution.

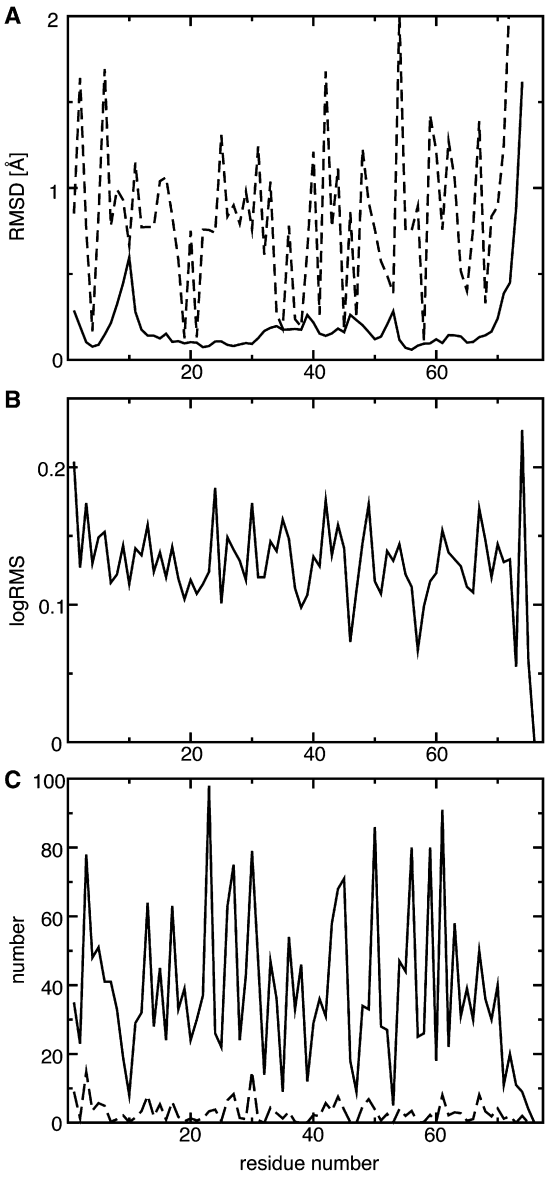

Figure 3. Structural Statistics as a Function of Residue for Ubiquitin

(A) RMS difference from the average structure, for backbone (solid line) and all non-hydrogen atoms (dashed line).

(B) Logarithmic RMS differences $\left(\sqrt{\sum_i log^2(d^i_{obs}/d^i_{calc})/n_k}\right)$ per residue k, where n_k is the number of residues involving residue k.

(C) Number of restraints n_k (solid line) and number of logarithmic discrepancies $log(d/d_0)$ larger than two σ (dashed line).

The local disorder is also apparent in Figure 4A–F, which shows structure ensembles of the 30% lowest energy structures for all examples. That local heterogeneity appears in regions with dynamic behavior that also applies to side chains in the hydrophobic core of the protein. Figure 4G shows the side chains analyzed in an ensemble refinement of Ubiquitin (Lindorff-Larsen et al., 2005). All residues showing major fluctuations in this ensemble refinement, with the exception of residue 13, also show heterogeneity in the current ensemble (residues 15, 44, and 67). This is an illustration of the multiple minima created by the properties of the log-harmonic potential for inconsistent data (see Figure 1).

Due to the large quantity and the quality of the distance restraints for Ubiquitin, we observe only a small reduction for the RMS difference to the X-ray crystal structure (Vijay-Kumar et al., 1987). For the average structures, the coordinate RMS difference reduces from 0.51 to 0.46 Å. For the other examples, the log–-harmonic potential has a larger impact on the RMS difference to the X-ray crystal structure, and sometimes results in substantial reduction in RMS difference for the average structures. A striking improvement was found for Il4, where the RMS distance between the average structure (calculated with hydrogen bonds and torsion angle restraints) and the X-ray crystal structure (Wlodawer, 1992) is 1.11 Å, compared to 1.74 Å for the originally deposited structures (Smith et al., 1994). For GB1, the RMS difference to the X-ray crystal structure (Gallagher et al., 1994) reduces from 1.14 to 0.55 Å, for BPTI (Marquart et al., 1983), it reduces from 0.78 to 0.59 Å, for the PH domain (Hyvonen et al., 1995), from 0.7 to 0.63 Å. For iL8 (Baldwin et al., 1991), there remain significant differences between NMR and X-ray crystal structure, in particular for the interhelical distance. These differences are expected, since some NOEs clearly indicate that the structures are different in solution and in the crystal. Nonetheless, the RMS difference reduces from around 2 Å to 1.6 Å, and we note that the interhelical distance is less well defined in the ensemble than other areas of the structure (e.g., the β sheet).

We also analyzed all structures with WhatIf (Hooft et al., 1996a) and ProCheck (Laskowski et al., 1993). Similar to what we observed previously with error-distribution derived potentials (Nilges et al., 2006), structures are, for most criteria, of better quality than those calculated with flat-bottom potentials. Some of the improvements were substantial (more than one or two standard deviations); for example, for IL4 (improvement of the packing Z-score QUACHK from −2.1 to −0.61, and of the Ramachandran quality Z-score RAMCHK from −4.4 to −2.6). Even for the much more complete Ubiquitin data set we observed an improvement of one standard deviation in the Ramachandran quality Z-score RAMCHK (from −2.9 to −1.8) (see Supplemental Data for a complete list). No distortions are present in the covalent geometry; the RMS Z-scores for bonds, bond angles, and planarity indicate very "tight" geometry. An analysis of Q-factors for Ubiquitin (Cornilescu et al., 1998) also shows a consistent improvement of the structures with the log-harmonic potential, sometimes by a substantial amount (e.g., from 0.524 to 0.346 for the N–H coupling; see Supplemental Data for a complete list).

CONCLUSION

We have presented a hybrid energy function that depends directly on the data quality and includes a distance restraint potential derived from a statistical analysis. We have developed an efficient minimization scheme for this hybrid energy. All terms in the restraint potential that depend on human bias (the width of bounds, the relative weight between force field and data) have been removed. The resulting structures are of better quality, as judged by independent validation criteria and their similarity to X-ray structures of the same molecule. Although the similarity to a X-ray crystal structure cannot be taken as a quality criterion by itself, it is an encouraging fact that the structures

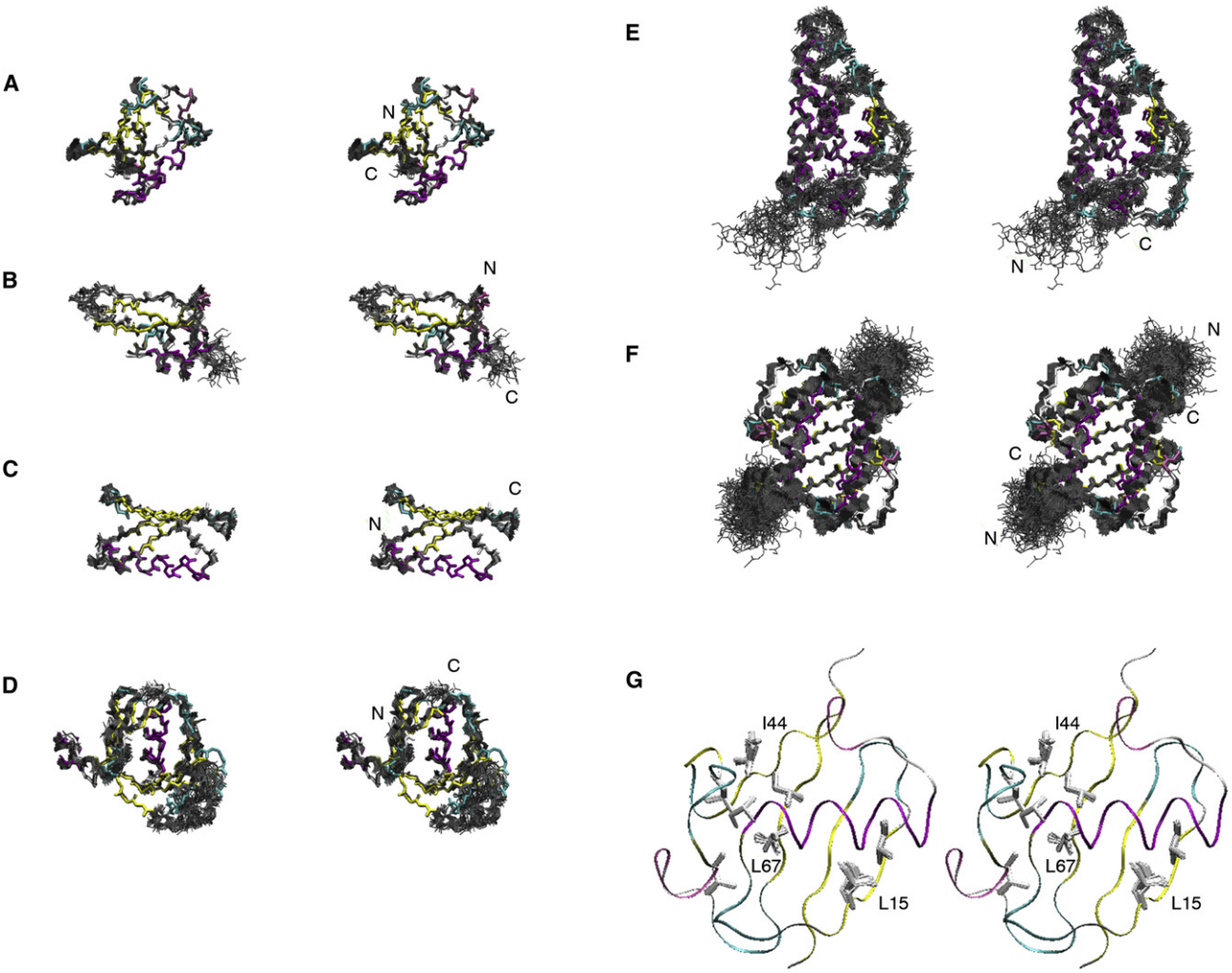

Figure 4. Structure Ensembles

(A–F) NMR structure ensembles superposed on the X–ray crystal structures. (A) GB1, (B) BPTI, (C) Ubiquitin, (D) β spectrin PH domain, (E) il4, and (F) il8. (G) Ubiquitin side chains in the hydrophobic core (Ile 13, Leu 15, Ile 23, Leu 43, Ile 44, Leu 50, Ile 61, Leu 67). The X–ray crystal structure is colored according to secondary structure (helices, magenta; β sheets, yellow; turns, cyan), and the NMR structures are colored gray. The Figure was prepared with VMD (Humphrey et al., 1996).

become more similar with the log-harmonic potential. No knowledge of the X-ray crystal structure is required in its derivation.

The approach described in this paper has some similarity to our entirely probabilistic calculation method, Inferential Structure Determination (ISD) (Rieping et al., 2005b). The principal difference is that ISD is not based on minimization, but on a much more exhaustive exploration of conformational space and a probabilistic interpretation of the results. For example, the data qualities are sampled over many values, and the result is a distribution, not a single value. Also, structure selection by ranking of hybrid energy is not necessary, since the sampling algorithm automatically samples more probable states more densely; the number of structures and parameters in a particular region of coordinate and parameter space is a direct measure of the probability.

Despite the success of ISD in obtaining unbiased estimates of values of coordinates and their uncertainties, standard structure determination by minimization will undoubtedly continue to play a major role in practical applications, mostly due to calculational efficiency. The approach presented in this paper captures the important features of ISD and, similar to ISD, it makes structure determination more objective. In contrast to ISD, some results of the calculation, in particular, the distribution of the structures around their average, will depend, however, on the properties of the minimization algorithm. Similar to ISD, a crucial parameter is determined automatically by the structure calculation, and the evaluation of the data quality is a result of the calculation.

The much higher computational costs in ISD are justified, since ISD gives objective and statistically valid estimates of the

precision of coordinates and other parameters, which a minimization method, even if it is based on rigorous statistical analysis, cannot. The present paper shows that the maximum benefit from a minimization approach can be obtained when it is based on a rigorous probabilistic treatment.

An important reason for the development of the new minimization approach is that it is much faster and better suited for integration into iterative assignment and structure calculation algorithms, such as ARIA (Rieping et al., 2007). We are in the process of incorporating the joint energy, $E_{joint}(X, \sigma)$, and the log–harmonic potential fully into ARIA, in order to replace the ad hoc criteria for the rejection of a restraint (based on the violation of a bound) by statistically more meaningful ones.

The increase in similarity to the corresponding X-ray structure has been achieved by a strategy that is the exact opposite of the recently proposed ROSETTA refinement (Qian et al., 2007), where structures were refined without the data. We feel that experimental structures should be principally determined from the available experimental data. Consequently, in the present approach, we have deliberately used a simple "geometric" force field (no electrostatics, but repulsive van der Waals energy only) to underline the improvements that can be obtained by data treatment. In contrast to the ROSETTA refinement, our refinement takes a little more time than a standard structure calculation. Obviously, it can be combined with more elaborate force fields and more extensive sampling schemes. The additional improvement of the structures due to refinement in explicit water (Linge et al., 2003) is documented in the Supplemental Data.

The log-harmonic potential and the joint energy, $E_{joint}(X, \sigma)$, will certainly be useful for other applications where distance restraints are used; for example, other NMR parameters, such as paramagnetic relaxation enhancement (see Gillespie and Shortle, 1997) or structure prediction and comparative modeling methods that are based on the satisfaction of spatial restraints (Sali and Blundell, 1993). In particular, the properties of the potential with conflicting restraints may be useful for modeling with distances derived from multiple templates.

SUPPLEMENTAL DATA

Supplemental data include two tables, one figure, Supplemental Experimental Procedures, and Supplemental References and can be found with this article online at http://www.structure.org/cgi/content/full/16/9/1305/DC1/.

ACKNOWLEDGMENTS

This work was funded by EC grants QLG2-CT-2000-01313 and QLG2–CT–2002–00988 (to M.N.) and the ACI IMPBio ICMD-RMN (to T.M.).

Received: May 13, 2008
Revised: July 11, 2008
Accepted: July 20, 2008
Published: September 9, 2008

REFERENCES

Baldwin, E.T., Weber, I.T., Charles, R.S., Xuan, J.C., Appella, E., Yamada, M., Matsushima, K., Edwards, B.F., Clore, G.M., Gronenborn, A.M., et al. (1991). Crystal structure of interleukin 8: symbiosis of NMR and crystallography. Proc. Natl. Acad. Sci. USA *88*, 502–506.

Berndt, K.D., Güntert, P., Orbons, L.P., and Wüthrich, K. (1992). Determination of a high-quality nuclear magnetic resonance solution structure of the bovine pancreatic trypsin inhibitor and comparison with three crystal structures. J. Mol. Biol. *227*, 757–775.

Brunger, A.T., and Nilges, M. (1993). Computational challenges for macromolecular structure determination by x-ray crystallography and solution NMR-spectroscopy. Q. Rev. Biophys. *26*, 49–125.

Brunger, A.T., Krukowski, A., and Erickson, J.W. (1990). Slow-cooling protocols for crystallographic refinement by simulated annealing. Acta Crystallogr. A. *46*, 585–593.

Brunger, A.T., Clore, G.M., Gronenborn, A.M., Saffrich, R., and Nilges, M. (1993). Assessing the quality of solution nuclear magnetic resonance structures by complete cross-validation. Science *261*, 328–331.

Brunger, A.T., Adams, P.D., Clore, G.M., DeLano, W.L., Gros, P., Grosse-Kunstleve, R.W., Jiang, J.S., Kuszewski, J., Nilges, M., Pannu, N.S., et al. (1998). Crystallography and NMR system: a new software suite for macromolecular structure determination. Acta Crystallogr. D Biol. Crystallogr. *54*, 905–921.

Clore, G.M., and Schwieters, C.D. (2004). How much backbone motion in ubiquitin is required to account for dipolar coupling data measured in multiple alignment media as assessed by independent cross-validation? J. Am. Chem. Soc. *126*, 2923–2938.

Clore, G.M., Appella, E., Yamada, M., Matsushima, K., and Gronenborn, A.M. (1990). Three-dimensional structure of interleukin 8 in solution. Biochemistry *29*, 1689–1696.

Cornilescu, G., Marquardt, J.L., Ottiger, M., and Bax, A. (1998). Validation of protein structure from anisotropic carbonyl chemical shifts in a dilute liquid crystalline phase. J. Am. Chem. Soc. *120*, 6836–6837.

Damm, K.L., and Carlson, H.A. (2006). Gaussian-weighted RMSD superposition of proteins: a structural comparison for flexible proteins and predicted protein structures. Biophys. J. *90*, 4558–4573.

Gallagher, T., Alexander, P., Bryan, P., and Gilliland, G.L. (1994). Two crystal structures of the b1 immunoglobulin-binding domain of streptococcal protein g and comparison with nmr. Biochemistry *33*, 4721–4729.

Gillespie, J.R., and Shortle, D. (1997). Characterization of long-range structure in the denatured state of staphylococcal nuclease. I. Paramagnetic relaxation enhancement by nitroxide spin labels. J. Mol. Biol. *268*, 158–169.

Gronenborn, A.M., Filpula, D.R., Essig, N.Z., Achari, A., Whitlow, M., Wingfield, P.T., and Clore, G.M. (1991). A novel, highly stable fold of the immunoglobulin binding domain of streptococcal protein g. Science *253*, 657–661.

Habeck, M., Rieping, W., and Nilges, M. (2006). Weighting of experimental evidence in macromolecular structure determination. Proc. Natl. Acad. Sci. USA *103*, 1756–1761.

Hooft, R.W., Vriend, G., Sander, C., and Abola, E.E. (1996). Errors in protein structures. Nature *381*, 272.

Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD: visual molecular dynamics. J. Mol. Graph. *14*, 33–38.

Hyvonen, M., Macias, M.J., Nilges, M., Oschkinat, H., Saraste, M., and Wilmanns, M. (1995). Structure of the binding site for inositol phosphates in a PH domain. EMBO J. *14*, 4676–4685.

Jack, A., and Levitt, M. (1978). Refinement of large structures by simultaneous minimization of energy and R factor. Acta Crystallogr. A *34*, 931–935.

Kaptein, R., Zuiderweg, E.R.P., Scheek, R., Boelens, R., and van Gunsteren, W.F. (1985). A protein structure from NMR data. lac repressor headpiece. J. Mol. Biol. *182*, 179–182.

Laskowski, R.A., MacArthur, M.W., Moss, D.S., and Thornton, J.M. (1993). PROCHECK: a program to check the stereochemical quality of protein structures. J. Appl. Cryst. *26*, 283–291.

Lindorff-Larsen, K., Best, R.B., Depristo, M.A., Dobson, C.M., and Vendruscolo, M. (2005). Simultaneous determination of protein structure and dynamics. Nature *433*, 128–132.

Linge, J.P., O'Donoghue, S.I., and Nilges, M. (2001). Automated assignment of ambiguous nuclear overhauser effects with aria. Methods Enzymol. *339*, 71–90.

Linge, J.P., Williams, M.A., Spronk, C.A.E.M., Bonvin, A.M.J.J., and Nilges, M. (2003). Refinement of protein structures in explicit solvent. Proteins *50*, 496–506.

Marquart, M., Walter, J., Deisenhofer, J., Bode, W., and Huber, R. (1983). The geometry of the reactive site and of the peptide groups in trypsin, trypsinogen and its complexes with inhibitors. Acta Crystallogr. B *39*, 480–485.

Nabuurs, S.B., Spronk, C.A.E.M., Krieger, E., Maassen, H., Vriend, G., and Vuister, G.W. (2003). Quantitative evaluation of experimental nmr restraints. J. Am. Chem. Soc. *125*, 12026–12034.

Nabuurs, S.B., Spronk, C.A., Vuister, G.W., and Vriend, G. (2006). Traditional biomolecular structure determination by nmr spectroscopy allows for major errors. PLoS Comput. Biol. *2*, e9.

Nilges, M., Clore, G.M., and Gronenborn, A.M. (1987). A simple method for delineating well-defined and variable regions in protein structures determined from interproton distance data. FEBS Lett. *219*, 11–16.

Nilges, M., Macias, M.J., O'Donoghue, S.I., and Oschkinat, H. (1997). Automated noesy interpretation with ambiguous distance restraints: the refined NMR solution structure of the pleckstrin homology domain from beta-spectrin. J. Mol. Biol. *269*, 408–422.

Nilges, M., Habeck, M., O'Donoghue, S.I., and Rieping, W. (2006). Error distribution derived distance potentials. Proteins *64*, 652–664.

Powers, R., Garrett, D.S., March, C.J., Frieden, E.A., Gronenborn, A.M., and Clore, G.M. (1992). Three-dimensional solution structure of human interleukin-4 by multidimensional heteronuclear magnetic resonance spectroscopy. Science *256*, 1673–1677.

Press, W., Flannery, B., Teukolsky, A., and Vetterling, W. (1986). Numerical Recipes: The Art of Scientific Computing (New York: Cambridge University Press).

Qian, B., Raman, S., Das, R., Bradley, P., McCoy, A.J., Read, R.J., and Baker, D. (2007). High-resolution structure prediction and the crystallographic phase problem. Nature *450*, 259–264.

Rieping, W. (2004). Quality Criteria for Protein NMR Structures (Regensburg, Germany: University of Regensburg).

Rieping, W., Habeck, M., and Nilges, M. (2005a). Inferential structure determination. Science *309*, 303–306.

Rieping, W., Habeck, M., and Nilges, M. (2005b). Modeling errors in noe data with a lognormal distribution improves the quality of nmr structures. J. Am. Chem. Soc. *127*, 16026–16027.

Rieping, W., Habeck, M., Bardiaux, B., Bernard, A., Malliavin, T.E., and Nilges, M. (2007). ARIA2: automated noe assignment and data integration in nmr structure calculation. Bioinformatics *23*, 381–382.

Sali, A., and Blundell, T.L. (1993). Comparative protein modeling by satisfaction of spatial restraints. J. Mol. Biol. *234*, 779–815.

Smith, L.J., Redfield, C., Smith, R.A., Dobson, C.M., Clore, G.M., Gronenborn, A.M., Walter, M.R., Naganbushan, T.L., and Wlodawer, A. (1994). Comparison of four independently determined structures of human recombinant interleukin-4. Nat. Struct. Biol. *1*, 301–310.

Vijay-Kumar, S., Bugg, C.E., and Cook, W.J. (1987). Structure of ubiquitin refined at 1.8 å resolution. J. Mol. Biol. *194*, 531–544.

Wlodawer, A. (1992). Crystal structure of human recombinant interleukin-4 at 2.25 å resolution. FEBS Lett. *309*, 59–64.

